



PATENT ABSTRACTS OF JAPAN

(11) Publication number: **09134360 A**(43) Date of publication of application: **20 . 05 . 97**

(51) Int. Cl.

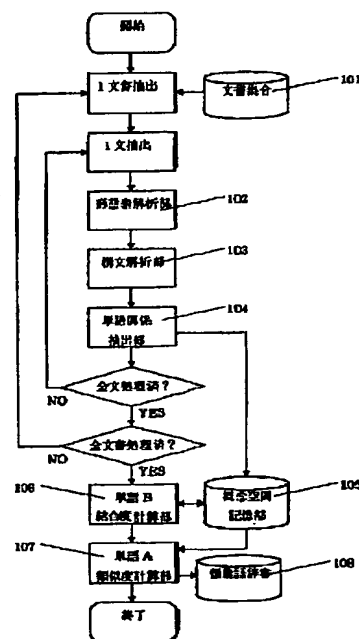
G06F 17/28(21) Application number: **07293062**(71) Applicant: **OMRON CORP**(22) Date of filing: **10 . 11 . 95**(72) Inventor: **FUJII FUJIKI**(54) **METHOD AND DEVICE FOR QUANTIFYING CONCEPT OF 'WORD' AND METHOD AND DEVICE UTILIZING THE SAME**

(57) Abstract:

PROBLEM TO BE SOLVED: To quantize the concept of words suitable for generating similarity between words.

SOLUTION: The method for quantifying the concept of 'word' used in a document is provided with a step 104 for extracting one or more than two 'relative words' having relation of forming a grammatical group with the 'word' by analyzing the applied document and a step 106 for finding out 'the degree of connection' between the 'word' and each of the 'relative words'. The concept of the 'word' is quantized by 'the degree of connection' to each of the 'relative words' having relation of forming the grammatical group with the 'word'.

COPYRIGHT: (C)1997,JPO



(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平9-134360

(43) 公開日 平成9年(1997)5月20日

(51) Int.Cl.⁶

G 0 6 F 17/28

識別記号

片内整理番号

F I

G 0 6 F 15/38

技術表示箇所

C

審査請求 未請求 請求項の数24 O L (全 16 頁)

(21) 出願番号 特願平7-293062

(22) 出願日 平成7年(1995)11月10日

(71) 出願人 000002945

オムロン株式会社

京都府京都市右京区花園土堂町10番地

(72) 発明者 藤居 藤樹

京都府京都市右京区花園土堂町10番地 オムロン株式会社内

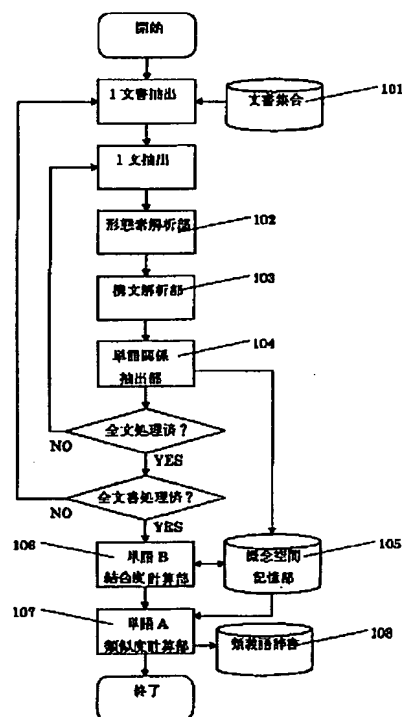
(74) 代理人 弁理士 飯塚 信市

(54) 【発明の名称】 『語』の概念を定量化するための方法及び装置、並びに、それらを用いた方法及び装置

(57) 【要約】

【課題】 語相互間の類似度生成に好適な、語の概念を定量化するための方法及び装置を提供すること。

【解決手段】 文書中で用いられた『語』の概念を定量化するための方法であって、与えられた文書を解析することにより、前記『語』と文法上の組を形成する関係にある1若しくは2以上の『関係語』を抽出するステップと、前記『語』が前記1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』を求めるステップと、を具備し、前記『語』の概念を、それと文法上の組を形成する関係にある1若しくは2以上の『関係語』のそれぞれに対する『結合度』の形で定量化する、ことを特徴とする。



【特許請求の範囲】

【請求項1】 文書中で用いられた『語』の概念を定量化するための方法であって、
与えられた文書を解析することにより、前記『語』と文法上の組を形成する関係にある1若しくは2以上の『関係語』を抽出するステップと、
前記『語』が前記1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』を求めるステップと、を具備し、
前記『語』の概念を、それと文法上の組を形成する関係にある1若しくは2以上の『関係語』のそれぞれに対する『結合度』の形で定量化する、ことを特徴とする方法。

【請求項2】 文書中で用いられた『語』相互間の類似度を生成するための方法であって、
与えられた文書を解析することにより、比較対象となる『語』のそれぞれについて、その『語』と文法上の組を形成する関係にある1若しくは2以上の『関係語』を抽出するステップと、
比較対象となる『語』のそれぞれについて、その『語』が前記1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』を求めて、各『語』の概念をそれと関係する『関係語』のそれぞれを座標軸としかつ各『関係語』との『結合度』のそれぞれを軸成分とする概念ベクトルとして生成するステップと、
比較対象となる『語』のそれぞれを、その概念ベクトル同士で相互に比較することにより『語』相互間の類似度を生成するステップと、
を具備する、ことを特徴とする方法。

【請求項3】 文書中で用いられた『語』から類義語辞書を構築するための方法であって、
与えられた文書を解析することにより、辞書化の対象となる『語』のそれぞれについて、その『語』と文法上の組を形成する関係にある1若しくは2以上の『関係語』を抽出するステップと、
辞書化の対象となる『語』のそれぞれについて、その『語』が前記1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』を求めて、各『語』の概念をそれと関係する『関係語』のそれぞれを座標軸としかつ各『関係語』との『結合度』のそれぞれを軸成分とする概念ベクトルとして生成するステップと、
辞書化の対象となる『語』のそれぞれを、その概念ベクトル同士で相互に比較することにより『語』相互間の類似度を生成するステップと、
前記類似度に基づいて類似すると判定される『語』同志を関連付けて類義語辞書を構築するステップと、
を具備する、ことを特徴とする方法。

【請求項4】 文書中で用いられた『語』から類義語辞書を構築するための方法であって、
与えられた文書を解析することにより、辞書化の対象と

なる『語』のそれぞれについて、その『語』と文法上の組を形成する関係にある1若しくは2以上の『関係語』を、複数の文法上の組のそれぞれについて抽出するステップと、

辞書化の対象となる『語』のそれぞれについて、その『語』が前記1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』を、複数の文法上の組のそれぞれについて求めて、各『語』の概念をそれと関係する『関係語』のそれぞれを座標軸としかつ各『関係語』との『結合度』のそれぞれを軸成分とする複数の概念ベクトルとして生成するステップと、
辞書化の対象となる『語』のそれぞれを、同一の文法上の組に対応して生成される概念ベクトル同士で相互に比較することにより『語』相互間の類似度を複数の文法上の組のそれぞれについて生成するステップと、
前記複数の文法上の組のそれぞれに対応して生成される複数の類似度に基づいて総合的に類似すると判定される『語』同志を関連付けて類義語辞書を構築するステップと、
を具備する、ことを特徴とする方法。

【請求項5】 前記文法上の組を形成する関係とは、動詞とその目的語の組を形成する関係である、ことを特徴とする請求項1乃至請求項4のいずれかに記載の方法。

【請求項6】 前記文法上の組を形成する関係とは、修飾語とその被修飾語の組を形成する関係である、ことを特徴とする請求項1乃至請求項4のいずれかに記載の方法。

【請求項7】 前記文法上の組を形成する関係とは、主語とその述語の組を形成する関係である、ことを特徴とする請求項1乃至請求項4のいずれかに記載の方法。

【請求項8】 前記『語』が1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』とは、着目している『語』に関する各『関係語』の出現回数のそれぞれを正規化した値である、ことを特徴とする請求項1乃至請求項4のいずれかに記載の方法。

【請求項9】 前記正規化した値とは、着目している『語』に関する全『関係語』の出現回数に対する個々の『関係語』の出現回数の割合である、ことを特徴とする請求項8に記載の方法。

【請求項10】 前記正規化した値とは、着目している『語』に関する各『関係語』の出現回数の中で最大の出現回数に対する個々の『関係語』の出現回数の割合である、ことを特徴とする請求項8に記載の方法。

【請求項11】 前記文書は、文書データベースから読み出されて与えられる、ことを特徴とする請求項1乃至請求項4のいずれかに記載の方法。

【請求項12】 前記文書は、通信回線を介して順次与えられる、ことを特徴とする請求項1乃至請求項4のいずれかに記載の方法。

【請求項13】 文書中で用いられた『語』の概念を定

量化するための装置であって、
与えられた文書を解析することにより、前記『語』と文法上の組を形成する関係にある1若しくは2以上の『関係語』を抽出する手段と、
前記『語』が前記1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』を求める手段と、を具備し、
前記『語』の概念を、それと文法上の組を形成する関係にある1若しくは2以上の『関係語』のそれぞれに対する『結合度』の形で定量化する、ことを特徴とする装置。

【請求項14】 文書中で用いられた『語』相互間の類似度を生成するための装置であって、
与えられた文書を解析することにより、比較対象となる『語』のそれぞれについて、その『語』と文法上の組を形成する関係にある1若しくは2以上の『関係語』を抽出する手段と、
比較対象となる『語』のそれぞれについて、その『語』が前記1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』を求めて、各『語』の概念をそれと関係する『関係語』のそれぞれを座標軸としかつ各『関係語』との『結合度』のそれぞれを軸成分とする概念ベクトルとして生成する手段と、
比較対象となる『語』のそれぞれを、その概念ベクトル同士で相互に比較することにより『語』相互間の類似度を生成する手段と、
を具備する、ことを特徴とする装置。

【請求項15】 文書中で用いられた『語』から類義語辞書を構築するための装置であって、
与えられた文書を解析することにより、辞書化の対象となる『語』のそれぞれについて、その『語』と文法上の組を形成する関係にある1若しくは2以上の『関係語』を抽出する手段と、
辞書化の対象となる『語』のそれぞれについて、その『語』が前記1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』を求めて、各『語』の概念をそれと関係する『関係語』のそれぞれを座標軸としかつ各『関係語』との『結合度』のそれぞれを軸成分とする概念ベクトルとして生成する手段と、辞書化の対象となる『語』のそれぞれを、その概念ベクトル同士で相互に比較することにより『語』相互間の類似度を生成する手段と、
前記類似度に基づいて類似すると判定される『語』同志を関連付けて類義語辞書を構築する手段と、
を具備する、ことを特徴とする装置。

【請求項16】 文書中で用いられた『語』から類義語辞書を構築するための装置であって、
与えられた文書を解析することにより、辞書化の対象となる『語』のそれぞれについて、その『語』と文法上の組を形成する関係にある1若しくは2以上の『関係語』

を、複数の文法上の組のそれぞれについて抽出する手段と、

辞書化の対象となる『語』のそれぞれについて、その『語』が前記1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』を、複数の文法上の組のそれぞれについて求めて、各『語』の概念をそれと関係する『関係語』のそれぞれを座標軸としかつ各『関係語』との『結合度』のそれぞれを軸成分とする複数の概念ベクトルとして生成する手段と、
辞書化の対象となる『語』のそれぞれを、同一の文法上の組に対応して生成される概念ベクトル同士で相互に比較することにより『語』相互間の類似度を複数の文法上の組のそれぞれについて生成する手段と、
前記複数の文法上の組のそれぞれに対応して生成される複数の類似度に基づいて総合的に類似すると判定される『語』同志を関連付けて類義語辞書を構築する手段と、
を具備する、ことを特徴とする装置。

【請求項17】 前記文法上の組を形成する関係とは、動詞とその目的語の組を形成する関係である、ことを特徴とする請求項13乃至請求項16のいずれかに記載の装置。

【請求項18】 前記文法上の組を形成する関係とは、修飾語とその被修飾語の組を形成する関係である、ことを特徴とする請求項13乃至請求項16のいずれかに記載の装置。

【請求項19】 前記文法上の組を形成する関係とは、主語とその述語の組を形成する関係である、ことを特徴とする請求項13乃至請求項16のいずれかに記載の装置。

【請求項20】 前記『語』が1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』とは、着目している『語』に関する各『関係語』の出現回数のそれぞれを正規化した値である、ことを特徴とする請求項13乃至請求項16のいずれかに記載の装置。

【請求項21】 前記正規化した値とは、着目している『語』に関する全『関係語』の出現回数に対する個々の『関係語』の出現回数の割合である、ことを特徴とする請求項20に記載の装置。

【請求項22】 前記正規化した値とは、着目している『語』に関する各『関係語』の出現回数の中で最大の出現回数に対する個々の『関係語』の出現回数の割合である、ことを特徴とする請求項20に記載の装置。

【請求項23】 前記文書は、文書データベースから読み出されて与えられる、ことを特徴とする請求項13乃至請求項16のいずれかに記載の装置。

【請求項24】 前記文書は、通信回線を介して順次与えられる、ことを特徴とする請求項13乃至請求項16のいずれかに記載の装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】この発明は、与えられた文書からそれに使用された『語』の概念を定量化するための方法及び装置、並びに、それらを用いて『語』相互の類似度を生成したり、或いは、類義語辞書を自動構築したりするための方法及び装置に関する。

【0002】

【従来の技術】文書検索の分野においては、同義語・類義語辞書を用いて利用者の検索数を拡張し、これにより検索に際するヒット率を上げるといった手法が広く採用されている。従来、この種の文書検索用辞書の構築は専ら人手により行われているため、その作業工数が膨大なことから、そのような辞書は非常に高価なものとなっている。そこで、昨今、この種の文書検索用辞書を低コストに構築するために、検索対象となる文書集合中から共起統計を用いて得られる共起ベクトルに基いて文書検索用辞書を計算機により自動構築する装置が提案されている。

【0003】

【発明が解決しようとする課題】しかしながら、このような文書検索用辞書の自動構築装置にあっては、共起頻度を算出する際に、単語の持つ文法的・意味的な性質についての考慮が払われていないため、自動構築される辞書の性能は必ずしも満足の得られるものではなく、結果としてそのようにして自動構築された辞書を用いた場合、文書検索に際するヒット率が低かった。

【0004】この発明は、このような従来の問題点に着目してなされたものであり、その目的とするところは、文書検索等の用途に適する高性能な類義語辞書を自動的に構築するための方法及び装置を提供することにある。この発明の他の目的とするところは、この種の辞書の自動構築に好適な、文書中に使用された語相互間の類似度を生成するための方法及び装置を提供することにある。この発明のさらに他の目的とするところは、この種の語相互間の類似度生成に好適な、語の概念を定量化するための方法及び装置を提供することにある。

【0005】

【課題を解決するための手段】この出願の請求項1（又は請求項13）に記載の発明は、文書中で用いられた『語』の概念を定量化するための方法（又は装置）であって、与えられた文書を解析することにより、前記『語』と文法上の組を形成する関係にある1若しくは2以上の『関係語』を抽出するステップ（又は手段）と、前記『語』が前記1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』を求めるステップ（又は手段）と、を具備し、前記『語』の概念を、それと文法上の組を形成する関係にある1若しくは2以上の『関係語』のそれぞれに対する『結合度』の形で定量化する、ことを特徴とする方法（又は装置）である。

【0006】ここで、『与えられた文書を解析する』とは、例えば、文書を構成する各単語を同定したり、品詞

を決定したり、更には、文の構成要素や修飾関係を同定すること等を意味している。尚、ここで言う『文書』には、日本語に限らず、後述する『文法上の組を形成する関係』が存在する限り、英語、フランス語、ドイツ語、中国語等々のあらゆる外国語が含まれる。

【0007】また、『文法上の組を形成する関係』とは、動詞と目的語との関係、修飾語と被修飾語との関係、主語と述語との関係、等のように、文法上において互いに密接に結合される関係のことを意味している。

【0008】また、『語』とはその概念を定量化すべく着目されている語のことであり、『関係語』とは上記の『語』に対して上記の『文法上の組を形成する関係』を有する語のことを意味している。例えば、上記の『関係』が動詞と目的語との関係である場合、動詞を『語』とすれば目的語が『関係語』となるし、逆に、目的語を『語』とすれば動詞が『関係語』となる。上記の関係が、修飾語と被修飾語との関係、主語と述語との関係である場合も、それぞれ同様である。

【0009】また、『結合度』とは、着目されている『語』がその『関係語』と同時に使用される度合いを表す数値のことを意味しており、具体的には、対象となる文書中において前記『語』が前記の『文法上の組を形成する関係』をもって前記『関係語』と同時に出現する回数やその回数を正規化した値により表現することができる。

【0010】そして、この請求項1（又は請求項13）に記載の発明によれば、着目されている『語』の概念は、それと文法上の組を形成する関係にある1若しくは2以上の『関係語』のそれぞれに対する『結合度』の形で定量化される。これは、着目されている『語』の概念が、各『関係語』をその座標軸としかつ各『結合度』を軸成分とする概念ベクトルとして表されたことを意味しており、別言すれば、着目されている『語』は、その『語』に固有な一群の用法の形で客観的に把握されることを意味している。

【0011】尚、定量化された語概念の精度を一定値以上に維持するためには、多義語（複数の意味を持つ語）をあらかじめ登録しておき、そのような多義語については着目される『語』の対象から排除することが好ましいであろう。

【0012】この出願の請求項2（又は請求項14）に記載の発明は、文書中で用いられた『語』相互間の類似度を生成するための方法（又は装置）であって、与えられた文書を解析することにより、比較対象となる『語』のそれぞれについて、その『語』と文法上の組を形成する関係にある1若しくは2以上の『関係語』を抽出するステップ（又は手段）と、比較対象となる『語』のそれぞれについて、その『語』が前記1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』を求めて、各『語』の概念をそれと関係する『関係語』のそれ

それを座標軸としそれらの『関係語』との『結合度』のそれぞれを軸成分とする概念ベクトルとして生成するステップ（又は手段）と、比較対象となる『語』のそれぞれを、その概念ベクトル同士で相互に比較することにより『語』相互間の類似度を生成するステップ（又は手段）と、を具備する、ことを特徴とする方法（又は装置）である。

【0013】ここで、『概念ベクトル同志の比較』は公知の数学的な種々の方法により行うことができ、例えば、比較されるべき2つの概念ベクトルのなす角度の余弦を求めたり、或いは、比較されるべき2つの概念ベクトルの距離を求めることにより、行うことができる。このとき、特に、後者の距離による場合には、正規化された『結合度』を用いることが好ましい。また、ここで言う『類似度』とは、2つの『語』が類似する度合いのことを意味し、その表現形態としては種々の形態を採用することができる。例えば、最も類似する場合を類似度『1.0』、最も類似しない場合を『0.0』とし、その間を連続的な少数により表現したり、あるいは最も類似する場合を『100%』、最も類似しない場合を『0%』とし、その間を連続的な百分率により表現することができる。さらに、別の表現形態としては、上記の少数や百分率を複数の閾値で弁別して、多段階の整数にて表現することもできる。

【0014】そして、この請求項2（又は請求項14）に記載の発明によれば、従来のように単に共起頻度のみ依存するのではなく、個々の『語』の有する概念を考慮した上で『語』相互間の類似度が生成され、その結果、人間の類似感覚に近い類似度が生成される。

【0015】この出願の請求項3（又は請求項15）に記載の発明は、文書中で用いられた『語』から類義語辞書を構築するための方法（又は装置）であって、与えられた文書を解析することにより、辞書化の対象となる『語』のそれぞれについて、その『語』と文法上の組を形成する関係にある1若しくは2以上の『関係語』を抽出するステップ（又は手段）と、辞書化の対象となる『語』のそれぞれについて、その『語』が前記1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』を求めて、各『語』の概念をそれと関係する『関係語』のそれぞれを座標軸としかつ各『関係語』との『結合度』のそれぞれを軸成分とする概念ベクトルとして生成するステップ（又は手段）と、辞書化の対象となる『語』のそれぞれを、その概念ベクトル同士で相互に比較することにより『語』相互間の類似度を生成するステップ（又は手段）と、前記類似度に基づいて類似すると判定される『語』同志を関連付けて類義語辞書を構築するステップ（又は手段）と、を具備する、ことを特徴とする方法（又は装置）である。

【0016】ここで、『類似すると判定される語同志を関連付けて』とは、ある語を指定するとそれと類似する

語が検索できることを意味しており、例えば、類似する語同志を一纏めにして登録したり、類似する語同志に類似度を付して類似する順に並べたり、共通のコードを付したりすることを意味している。

【0017】そして、この請求項3（又は請求項15）に記載の発明によれば、文書検索に適しかつ高性能な類義語辞書が自動的に構築される。

【0018】この出願の請求項4（又は請求項16）に記載の発明は、文書中で用いられた『語』から類義語辞書を構築するための方法（又は装置）であって、与えられた文書を解析することにより、辞書化の対象となる『語』のそれぞれについて、その『語』と文法上の組を形成する関係にある1若しくは2以上の『関係語』を、複数の文法上の組のそれぞれについて抽出するステップ（又は手段）と、辞書化の対象となる『語』のそれぞれについて、その『語』が前記1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』を、複数の文法上の組のそれぞれについて求めて、各『語』の概念をそれと関係する『関係語』のそれぞれを座標軸としかつ各『関係語』との『結合度』のそれぞれを軸成分とする複数の概念ベクトルとして生成するステップ（又は手段）と、辞書化の対象となる『語』のそれぞれを、同一の文法上の組に対応して生成される概念ベクトル同士で相互に比較することにより『語』相互間の類似度を複数の文法上の組のそれぞれについて生成するステップ（又は手段）と、前記複数の文法上の組のそれぞれに対応して生成される複数の類似度に基づいて総合的に類似すると判定される『語』同志を関連付けて類義語辞書を構築するステップ（又は手段）と、を具備する、ことを特徴とする方法（又は装置）である。

【0019】ここで、『複数の文法上の組のそれぞれについて』とは、例えば、『語』が目的語で『関係語』が動詞である場合だけではなく、『語』が被修飾語で『関係語』が修飾語である場合や、『語』が主語で『関係語』が述語である場合のように、同一の『語』を様々な『文法上の組』について、の意味である。

【0020】そして、この請求項4（又は請求項16）に記載の発明によれば、特定の文法上の組に偏ることなく、種々の文法上の組を考慮して、換言すれば、種々の文法上の用法を考慮して、類義語辞書の構築が行われるため、一層高性能な辞書の構築が可能となる。

【0021】この出願の請求項5（又は請求項17）に記載の発明は、前記請求項1（又は請求項13）乃至請求項4（又は請求項16）のいずれかに記載の方法（又は装置）において、前記文法上の組を形成する関係とは、動詞とその目的語の組を形成する関係である、ことを特徴とするものである。

【0022】ここで、『動詞とその目的語の組を形成する関係』とは、着目している『語』が動詞でその『関係語』が目的語の場合（前者）と、着目している『語』が

目的語でその『関係語』が動詞の場合（後者）との双方の場合を含む意味である。前者の場合には、着目された動詞に相当する『語』の概念は、それが目的語としてどのような『関係語』と強く結合するかと言った観点で定量化され、また後者の場合には、着目された目的語に相当する『語』の概念は、それが動詞としてどのような『関係語』と強く結合するかと言った観点で定量化される。

【0023】この出願の請求項6（又は請求項18）に記載の発明は、前記請求項1（又は請求項13）乃至請求項4（又は請求項16）のいずれかに記載の方法（又は装置）において、前記文法上の組を形成する関係とは、修飾語とその被修飾語の組を形成する関係である、ことを特徴とするものである。

【0024】ここで、『修飾語とその被修飾語の組を形成する関係』とは、着目している『語』が修飾語でその『関係語』が被修飾語の場合（前者）と、着目している『語』が被修飾語でその『関係語』が修飾語の場合（後者）との双方の場合を含む意味である。前者の場合には、着目された修飾語に相当する『語』の概念は、それが被修飾語としてどのような『関係語』と強く結合するかと言った観点で定量化され、また後者の場合には、着目された被修飾語に相当する『語』の概念は、それが修飾語としてどのような『関係語』と強く結合するかと言った観点で定量化される。

【0025】この出願の請求項7（又は請求項19）に記載の発明は、前記請求項1（又は請求項13）乃至請求項4（又は請求項16）のいずれかに記載の方法（又は装置）において、前記文法上の組を形成する関係とは、主語とその述語の組を形成する関係である、ことを特徴とするものである。

【0026】ここで、『主語とその述語の組を形成する関係』とは、着目している『語』が主語でその『関係語』が述語の場合（前者）と、着目している『語』が述語でその『関係語』が主語の場合（後者）との双方の場合を含む意味である。前者の場合には、着目された主語に相当する『語』の概念は、それが述語としてどのような『関係語』と強く結合するかと言った観点で定量化され、また後者の場合には、着目された述語に相当する『語』の概念は、それが主語としてどのような『関係語』と強く結合するかと言った観点で定量化される。

【0027】この出願の請求項8（又は請求項20）に記載の発明は、前記請求項1（又は請求項13）乃至請求項4（又は請求項16）のいずれかに記載の方法（又は装置）において、前記『語』が1若しくは2以上の『関係語』のそれぞれに対して有する『結合度』とは、着目している『語』に関する各『関係語』の出現回数のそれぞれを正規化した値である、ことを特徴とするものである。

【0028】『結合度』としてどのような値を使用すべ

きかは、その後、概念ベクトル同士を比較して類似度を求める際にどのような演算手法を用いるかに掛かっている。概念ベクトル同士の比較にベクトル同士のなす角度の余弦を用いるのであれば、『結合度』としては上記の出現回数それ自体を使用することができる。これに対して、概念ベクトル同士の比較にベクトル間の距離

（『語』の概念を各『関係語』を座標軸とする概念空間上の点としてとらえた場合において、そのような2点間の距離を求めるの意味）を用いるのであれば、『結合度』としては上記の出現回数を正規化したものを使用することが、類似度算出における精度向上のためには好ましい。

【0029】この出願の請求項9（又は請求項21）に記載の発明は、前記請求項8（又は請求項20）に記載の方法（又は装置）において、前記正規化した値とは、着目している『語』に関する全『関係語』の出現回数に対する個々の『関係語』の出現回数の割合である、ことを特徴とするものである。

【0030】この出願の請求項10（又は請求項22）に記載の発明は、前記請求項8（又は請求項20）に記載の方法（又は装置）において、前記正規化した値とは、着目している『語』に関する各『関係語』の出現回数の中で最大の出現回数に対する個々の『関係語』の出現回数の割合である、ことを特徴とするものである。

【0031】この出願の請求項11（又は請求項23）に記載の発明は、前記請求項1（又は請求項13）乃至請求項4（又は請求項16）のいずれかに記載の方法（又は装置）において、前記文書は、文書データベースから読み出されて与えられる、ことを特徴とする。

【0032】この発明において、精度の高い語概念定量化のためには大量の文書が必要であり、一般的には、このような大量の文書は磁気ディスクや光ディスク等の記録媒体に文書データベースとして格納されている。そして、この請求項では、このような文書データベースから順次読み出された文書データが語概念定量化処理の対象となり、その結果として得られる類義語辞書は当該文書データベースを対象とした文書検索に最適なものとなる。

【0033】この出願の請求項12（又は請求項24）に記載の発明は、前記請求項1（又は請求項13）乃至請求項4（又は請求項16）のいずれかに記載の方法（又は装置）において、前記文書は、通信回線を介して順次に与えられる、ことを特徴とするものである。

【0034】この発明において、精度の高い語概念定量化のためには大量の文書が必要であることは上述の通りである。この請求項では、例えば、インターネット等の通信回線を介して次々と送られてくる文書データが語概念定量化処理の対象となり、その結果として得られる類義語辞書はインターネット等を介してアクセスされる文書データベースを対象とした文書検索に最適なものとなる。

る。

【0035】

【発明の実施の形態】以下に、本発明の好適な実施の形態を添付図面を参照しながら詳細に説明する。先ず、本発明の概要を身近な例を挙げてわかりやすく説明する。名詞は、ある特定の動作を表す語（動詞）の集合を用いてその意味を表現することができると思われる。例えば、『コーヒー』と言う名詞は、『飲む』、『入れる』、『沸かす』、『買う』等の動詞を用いて表現することができる。同様に、『お茶』と言う名詞も、『飲む』、『入れる』、『沸かす』、『買う』等の動詞を用いて表現することができる。もっとも、例えば、『沸かす』と言う動詞に着目して、名詞『コーヒー』と共に使用される頻度と名詞『お茶』と共に使用される頻度とを比較すると、僅かではあるものの、名詞『コーヒー』と共に使用される頻度の方が高いものと思われる。

【0036】このように、任意の名詞は特定の動詞集合を用いてその意味を表現することができ、さらに、仮に、同一の動詞集合を用いて表現される名詞同士であっても、動詞集合を構成する個々の動詞に着目すると、その名詞の使用頻度（『結合度』）には各名詞固有の値が存在することが認められる。

【0037】従って、任意の名詞は、特定の『動詞集合』と、その名詞と個々の動詞との『結合度』とにより、ある程度の微妙なニュアンスまでも含めて定量化することができるはずである。

【0038】このような仮定は、名詞と動詞との関係に限られるものではなく、名詞と形容詞との関係や動詞と副詞との関係等にも広く当て嵌まる。例えば、『コーヒー』と言う名詞は、『ほろ苦い』、『香りの高い』、『甘い』、『熱い』、『冷たい』、『こはく色の』等の形容詞集合を用いてある程度は表現することができ、『跳ぶ』と言う動詞は、『高く』、『遠くへ』、『軽やかに』、『リズムカルに』等の副詞集合を用いてある程度は表現することができる。

【0039】さらに、上述した名詞と動詞との関係、名詞と形容詞との関係、動詞と副詞との関係は、それぞれ関係における品詞を相互に入れ替えた場合にもある程度は成立することが認められる。例えば、『飲む』と言う動詞は、『コーヒー』、『ジュース』、『酒』、『スープ』等の名詞集合によりある程度は表現することができ、『リズムカルに』と言う副詞は、『跳ねる』、『踊る』、『歌う』、『歩く』等の副詞によりある程度は表現することができる。

【0040】以上の仮定をより一般化すると、特定の品詞は、それと一定の関係にある他の品詞の集合、及び、個々の他の品詞との『結合度』を用いて定量的に表現できると言うことができる。

【0041】ところで、上述した相互に一定の関係にある、『品詞』と『他の品詞』とは、任意の文章中におい

ては、例えば、名詞と動詞との関係については、文法上の構文要素としての（主語と述語との組）を形成する関係や（述語動詞とその目的語との組）を形成する関係として、また動詞と副詞との関係や形容詞と名詞との関係については、文法上の構文要素としての（修飾語と被修飾語の組）を形成する関係として、それぞれ出現する。

【0042】従って、任意の文書集合が与えられた場合において、それに含まれる特定の『語』に着目し、その『語』と文法上の組（主語と述語との組、述語動詞とその目的語との組、修飾語と被修飾語の組等）を形成する関係にある『関係語』を抽出し、それら抽出された個々の『関係語』と着目されている『語』との『結合度』を求めれば、それら求められた各『関係語』毎の『結合度』を用いて、着目されている『語』の概念を定量化することができる。

【0043】ここで言う『語の概念』とは、言語学上一般に定義される『語の概念』とは若干異なる場合も生ずるであろう。先の例で言えば、本発明により生成される『語の概念』を用いた場合、『コーヒー』と『お茶』とは概念が類似するものと判断されるであろうが、果たして、言語学上においても両者が所謂『類義語』に相当するかについては、意見の分かれるところであろう。もっとも、このような言語学上の『語の概念』との相違は、文書中に生ずる語相互の比較や文書検索用類義語辞書の自動構築等の用途においては、さほど、問題とはならないであろう。

【0044】すなわち、本発明により生成される『語の概念』は、言わば、着目した『語』の文書集合中における『用法』を数値化したに過ぎないものではあるが、一方、言語学上において相互に類似するとされる複数の『語』についても、多くの場合、その文書中における『用法』は同様であろうから、本発明により生成される『語の概念』を使用して、文書中に生ずる語相互の比較や文書検索用類義語辞書の自動構築等を行っても、言語学上の『語の概念』を用いた場合と大幅な相違は生じないであろう。むしろ、本発明により生成される『語の概念』を使用した場合には、言語学上の『語の概念』よりも一層広い観点から語相互の比較を行いつつ文書検索用類義語辞書を自動構築することができる。そのため、このようにして自動構築された類義語辞書を用いて文書検索を行えば、従来の人間（言語学者）の主観を交えながら手作業で構築された類義語辞書を用いた場合や、単に共起頻度のみに基いて自動構築された類義語辞書を用いた場合に比較して、検索に際するヒット率を上げることができるであろう。

【0045】加えて、本発明による語相互の比較は、着目した『語』の文書集合中における『用法』を数値化したに過ぎない『語の概念』に基いて行われるものであって、その『語の概念』を既存の辞書に立ち戻って理解した上で行われるものではないから、対象となる文書中に

出現する『語』が新規に定義された技術用語であったり、更には、対象となる文書それ自体が不慣れな外国語であったとしても、その文書の構造が解析できる限り、支障なく語相互の比較を自動的に行うことができる。

【0046】従って、本発明によれば、単に、語の相互比較や類義語辞書の自動構築等の用途に止まらず、作者の異なる複数の文書集合の相互において、特定の『語』についての類似度比較を行うことにより、作者相互のものの考え方の相違を定量化して明らかとしたり、或いは、言語の異なる複数の文書集合の相互において、特定の『語』についての類似度比較を行うことにより、比較人類学的な考察を行う等のような幅広い応用が期待される。

【0047】次に、本発明の一実施形態である類義語辞書自動構築装置の構成を添付図面を参照しながら詳細に説明する。尚、本発明の実施のためには、公知の計算機システムが使用されるが、それらシステムのハードウェア構成については種々の文献により広く知られているため、その説明は省略する。さらに、以下に説明する装置には、同時に、本発明方法が含まれていることは当業者であれば容易に理解されるであろう。

【0048】本発明に係る類義語辞書自動構築装置のソフトウェア構成を示すネラルフローチャートを図1に示す。同図において、文書集合101は、類義語辞書を構築するための情報源として利用されるものであり、ここでは2つの意味を有している。すなわち、この類義語辞書自動構築装置が、文書検索用の類義語辞書を構築するために用いられる場合（前者）には、文書集合101はその検索対象となる文書集合それ自体を示している。他方、この類義語辞書自動構築装置が、インターネット上に存在する文書や電子メール等のような、通信回線を通じて入手される文書を類義語辞書を構築するための情報源として利用する場合（後者）には、そのようにして入手される文書それ自体を示している。

【0049】以下、前者の場合を前提として、説明を進めることとする。尚、後者の場合には、以下の処理は文書の入手に従い随時に実行されることとなる。文書集合101からは、それを構成する複数の文書が所定の順序にて1文書ずつ抽出され、さらに、その抽出された1文書を構成する複数の文が所定の順序で1文ずつ抽出されるようになっている。そして、以上の1文書抽出処理及び1文抽出処理は、文書集合を構成する全文書が抽出されるまで繰り返し行われる。このようにして順次に抽出される各文は、後述するように、形態素解析部102、構文解析部103、及び単語関係抽出部104において、形態素解析処理、構文解析処理、及び単語関係抽出処理に供される。

【0050】形態素解析部102は、抽出された1文に関して、形態素解析処理により単語の同定や品詞の決定を行う。この種の形態素解析処理は既に自然言語処理技

術として広く一般化されており、例えば、確率モデルを用いた方法やルールベースの方法等が知られている。

【0051】構文解析部103は、形態素解析部102にて抽出された形態素情報を元に文書中の各文から主語、動詞等の文の構成要素や修飾関係等を同定する。この種の構文解析処理も、既に、自然言語処理技術として一般化されている。

【0052】単語関係抽出部104は、構文解析部103にて抽出された構文情報を元に文書中の各文について単語関係に着目し、文法上の組を形成する関係にある2種類の構文要素（単語A、単語B）のそれぞれに相当する2個の単語を抽出する。尚、この抽出に際しては、後述する『結合度』を算出するための前処理として、特定の2個の単語の組毎に、それまでの文書中に出現した回数を計算して記憶しておく。

【0053】ここで、『文法上の組を形成する関係』としては、この例では、（目的語『単語A』，動詞『単語B』）の関係が選択されている。また、この例では、本出願の各請求項に言う『語』が（単語A）に相当する語に、また『関係語』が（単語B）に相当する語に、それぞれ対応している。

【0054】尚、『文法上の組を形成する関係』としては、それ以外にも、（主語『単語A』，述語『単語B』）の関係、（被修飾語『単語A』，修飾語『単語B』）の関係、更には、それらを逆にした、（動詞『単語A』，目的語『単語B』）の関係、（述語『単語A』，主語『単語B』）の関係、（修飾語『単語A』，被修飾語『単語B』）の関係等が、必要により適宜に選択可能になされている。このようにして抽出された2個の単語は、後述する、概念空間記憶部105の該当記憶エリアへと記憶される。

【0055】尚、最終的に構築される類義語辞書の性能を一層高めるためには、上述した構文解析部103と単語関係抽出部104との間に、一般にストップワード処理等と称される不要語削除処理やステミング等と称される語尾除去処理を設けることが好ましいと思われる。

【0056】上述した2個の単語が記憶される概念空間記憶部105の構造を図2に示す。同図に示されるように、概念空間記憶部105の構造は、特定の2個の構文要素（単語A、単語B）の組毎に、後述する、単語B結合度計算部106にて計算された『結合度』が記憶されるようになっている。また、ここで言う『概念空間』とは各単語Bを座標軸とする空間である。従って、文書集合101の中に出現する単語Bの種類数だけ座標軸が存在することとなる。そして、単語A（着目されている『語』）は、それと結合された各単語B（『関係語』）の『結合度』を軸成分とする座標値（又はベクトル）として概念空間上に配置される。尚、このとき、図2の概念空間記憶部105内に存在しない座標軸（単語B）は、『結合度』が0と言う意味であるから、その軸成分

は0とする。

【0057】単語B結合度計算部106は、概念空間記憶部105の内容に基いて、単語Bに相当する単語に対する単語Aに相当する単語の『結合度』を算出する。

尚、この単語B結合度計算部106における処理は、後述する単語A類似度計算部107における類似度計算数が『結合度』の正規化を必要とする場合にのみ必要なものであり、それ以外の場合には省略することが可能である。

【0058】単語B結合度計算部106における処理に関する詳細フローチャートを図3に示す。尚、同図において、 i 、 j はそれぞれ、単語A、単語Bのそれぞれに相当する語を順次に処理していくために用いる添字である。ここで、概念空間記憶部105に記憶された一連の単語Aは、一番目のものから i 番目のものまで、 A_1 （例えば、coffee）、 A_2 （例えば、tea）、 A_3 （例えば、water）、… A_i の如くに表される。他方、 i 番目の単語Aである A_i に結合された一連の単語Bは、一番目のものから j 番目のものまで、 B_{i1} 、 B_{i2} 、 B_{i3} 、… B_{ij} の如くに表される。

【0059】前述したように、単語関係抽出部104では、単語A及び単語Bに相当する特定の2個の単語の組毎に、それまでの文書中に出現した回数（単語Aに相当する特定の単語が出現した状態において、単語Bに相当する特定の単語が出現した回数）を計算して記憶している。

【0060】単語 B_{ij} 結合度計算部301では、単語 A_i が出現した状態において単語 B_{ij} が出現した出現回数を正規化することにより、各単語 B_{ij} 毎の『結合度』を計算する。出現回数を正規化して『結合度』を求めるための計算方法としては幾つかの方法が考えられる。第1の方法としては、着目している単語（ A_i ）に結合された全単語（ $B_{i1} \sim B_{in}$ ）の出現回数総和に対する個々の単語（ B_{i1} ）、（ B_{i2} ）、…（ B_{in} ）の出現回数の割合を求めるものである。第2の方法としては、着目している単語（ A_i ）に関する各単語（ B_{i1} ）、（ B_{i2} ）、…（ B_{in} ）の出現回数の中で最大の出現回数に対する個々の単語（ B_{i1} ）、（ B_{i2} ）、…（ B_{in} ）の出現回数の割合を求めるものである。尚、これらの計算方法については、後に、具体的な例を挙げて更に詳細に説明する。

【0061】以上までの処理にて、単語A（目的語）に相当する各単語（ A_i ）は、単語A（目的語）と文法上の組を形成する関係にある単語B（動詞）に相当する複数の単語（ B_{ij} ）、及び各単語（ B_{ij} ）に対する単語（ A_i ）の『結合度』の組（ A_i 、 B_{ij} 、『結合度』）により定量化されたことになる。尚、ここで、『結合度』については、前述したように、出現回数をそのまま用いる場合と、出現回数を正規化した値を用いる場合との2通りがある。

【0062】単語A類似度計算部107では、概念空間記憶部105に記憶された内容を元に、単語A（この例では目的語）の中の各単語同士の類似度を計算する。ここで、単語Aの中の各単語同士の類似度を計算する方法としては、2通りの方法が考えられる。第1の方法としては、単語Aの中の各単語が概念空間上のベクトル（概念ベクトル）として表現されていることに着目して、それらベクトル相互のなす角度の余弦を計算し、その計算結果を『類似度』とするものである。第2の方法としては、単語Aの中の各単語が概念空間上の点として表現されることに着目し、それら点間の距離を計算し、その計算結果を類似度とするものである。尚、これらの計算方法についても、後に、具体的な例を挙げて更に詳細に説明する。最後に、単語A類似度計算部107では、求められた単語Aの中の各単語相互の類似度に基づいて類義語辞書108を構築する。

【0063】このようにして構築された類義語辞書108の構造を図4に示す。同図に示されるように、類義語辞書108内には、各『単語』毎にそれに類似する『類義語』とその『類似度』とが対となって記憶されている。従って、この類義語辞書によれば、各『単語』の類義語領域及び類似度領域を参照することにより、その『単語』がどのような単語とどの程度に類似しているかを直ちに判断することができる。尚、ここで示された類義語辞書の構造は単なる一例に過ぎないものであり、その他、求められた類似度を適当な閾値にて弁別して多段階に表現したり、或いは各単語毎に類似する類義語を一纏めに記憶したり等の適宜な変形が可能なことは当業者であれば容易に理解されるであろう。

【0064】次に、本発明に係る類義語辞書自動構築装置の第2の実施形態を添付図面を参照して詳細に説明する。第2実施形態における類義語辞書自動構築装置のゼネラルフローチャートを図5に示す。尚、同図において、先の実施形態における図1のゼネラルフローチャートと同一構成部分については、同符号を付して詳細説明は省略する。

【0065】先の実施形態にて説明した『文法上の組を形成する関係』の中で、（目的語『単語A』、動詞『単語B』）の関係、（主語『単語A』、述語『単語B』）の関係、（被修飾語『単語A』、修飾語『単語B』）の関係をを用いて行われる類似度算出処理は、いずれも『名詞』若しくは『代名詞』同士の類似度を算出するものである。すなわち、（目的語『単語A』、動詞『単語B』）の関係、若しくは（主語『単語A』、述語『単語B』）の関係をを用いた類似度算出においては、動詞空間上において、『名詞』若しくは『代名詞』同士の類似度比較が行われる。また、（被修飾語『単語A』、修飾語『単語B』）の関係をを用いた類似度算出においては、形容詞空間上において、『名詞』若しくは『代名詞』同士の類似度比較が行われる。従って、比較されるべき『名

詞』若しくは『代名詞』同士が共通であるならば、それら3種類の関係から算出される3種類の類似度は本来一つに統合されなければならない。同様のことは、(動詞『単語A』, 目的語『単語B』)の関係、(動詞『単語A』, 主語『単語B』)の関係をj用いて算出される動詞同士の類似度の間でも言えるものである。

【0066】そこで、この第2の実施形態においては、先の実施形態において、複数種の『文法上の組を形成する関係』のそれぞれについて類義語辞書を構築した後、それら構築された複数種の類義語辞書を統合することにより、より精度の高い統合類義語辞書を構築するようにしている。

【0067】図5において、ステップ101~107は先の実施形態における同様の処理を行うものであり、これらの処理は『文法上の組を形成する関係』の全てが利用済みとなるまで(ステップ501)、繰り返行われる。尚、前述したように、ここで言う全ての『文法上の組を形成する関係』とは、(目的語『単語A』, 動詞『単語B』)の関係、(主語『単語A』, 述語『単語B』)の関係、(被修飾語『単語A』, 修飾語『単語B』)の関係、更には、それらを逆にした、(動詞『単語A』, 目的語『単語B』)の関係、(述語『単語A』, 主語『単語B』)の関係、(修飾語『単語A』, 被修飾語『単語B』)等が含まれる。その結果、類義語辞書108内には、『文法上の組を形成する関係』のそれぞれに対応する複数種の類義語辞書が構築される。

【0068】次いで、単語A類似度計算部502では、類義語辞書108内に構築された複数の類義語辞書を適宜に組み合わせることにより、単一の類似度を算出する。このとき、単一の類似度の算出は、先ず、個々の類義語辞書の中で、単語と類義語との組が同一のもの同士を組み合わせることにより行われる。次に、複数の類義語辞書に関しては、同一品詞についての類義語辞書同士を組み合わせることにより行われる。

【0069】前述したように、『文法上の組を形成する関係』の中で、(目的語『単語A』, 動詞『単語B』)の関係、(主語『単語A』, 述語『単語B』)の関係、(被修飾語『単語A』, 修飾語『単語B』)の関係をj用いて行われる類似度算出処理は、いずれも『名詞』若しくは『代名詞』同士の類似度を算出するものである。従って、比較されるべき『名詞』若しくは『代名詞』同士が共通であるならば、それら3種類の関係から算出され

I drink a cup of coffee. ... 文(1)

(動詞)

このとき、文法上の組を形成する関係として、目的語と動詞の組を形成する関係が設定されていると、文(1)に対して形態素解析処理、構文解析処理、単語関係抽出処理が施された結果として、目的語(coffee)と動詞(drink)とが抽出され、これらの単語は概念空間記憶部105内の該当領域に記憶され、同時に、そ

る3種類の類似度は本来一つに統合されなければならない。同様のことは、(動詞『単語A』, 目的語『単語B』)の関係、(動詞『単語A』, 主語『単語B』)の関係をj用いて算出される動詞同士の類似度の間でも言えるものである。

【0070】そこで、このような関係にある複数種の辞書については、単語A統合類似度計算部502の作用により互いに組み合わせられ、単語と類義語との組を共有する複数種の類似度については一つの類似度に統合される。統合の際の組み合わせ演算については、平均化演算、乗算、最大値演算、最小値演算等の各種の演算が利用される。単語A統合類似度計算部502の処理は、全品詞についての処理が完了するまで繰り返される(ステップ504)。尚、ここで言う『品詞』とは、対象となる文書集合が英語で記述されたものである場合、英語文法上の定義による11品種ではなく、文法的性質から大まかに区分されたものが使用されており、例えば、『名詞』と『代名詞』とは同一の『品詞』として取り扱われる。そして、この単語A統合類似度計算部502による演算結果を元に統合類義語辞書503が構築される。この統合類義語辞書は、類義語辞書108と同一の構造を有するものであるが、単語と類義語との組のそれぞれは2以上の類似度を有しない点で相違する。

【0071】このようにして構築される統合類義語辞書にあっては、語相互の類似度を複数の概念空間において比較して得られたものであるため、上述の組み合わせ演算を適切に設計することにより、語相互の類似関係を一層正確に反映したものとなり、これを文書検索等の用途に利用すれば、検索に際するヒット率を向上させることができる。

【0072】

【実施例】次に、本発明に係る類義語辞書自動構築装置の更に具体的な一実施例を添付図面を参照して詳細に説明する。尚、以下の例は、説明の便宜上、本発明に係る装置を英文で作成された文書集合に適用したが、日本語で作成された文書集合にも適用できることは勿論である。また、この例では、概念空間上のベクトル相互のなす角度の余弦を計算することにより、語相互の類似度が求められている。

【0073】今仮に、文書集合101より抽出された1文が、文(1)で示されるものであると想定する。

I drink a cup of coffee. ... 文(1)
(目的語)

の組についての出現回数のカウントアップが行われる。

【0074】同様の処理を繰り返しつつ、文書集合を構成する全ての文書の全ての文につき、形態素解析処理、構文解析処理、単語関係抽出処理が施された結果として、最終的に得られた概念空間記憶部の内容の一部を図6に示す。尚、この例では、『結合度』としては、目的

語（例えば、coffee）が出現した状態にて、動詞（例えば、drink）が出現した出現回数そのものが使用されており、前述した単語B結合度計算部106による正規化処理は行われていない。

【0075】図6から明らかなように、この例では、着目された単語A（目的語）である『coffee』は、3個の単語B（動詞）である『drink』、『boil』、『buy』と結合されており、それらの『結合度』はそれぞれ『10』、『4』、『1』とされている。同様にして、着目された単語A（目的語）である『tea』は、3個の単語B（動詞）である『drink』、『boil』、『buy』と結合されており、それらの『結合度』はそれぞれ『8』、『3』、『2』とされている。

【0076】図6においてその概念が定量化された2つ

の単語A（目的語）である『coffee』と『tea』とを概念空間上のベクトルとして表した状態を図7に示す。同図に示されるように、2つの単語A（目的語）である『coffee』と『tea』とは、3個の単語B（動詞）である『drink』、『boil』、『buy』をそれぞれ座標軸とし、かつそれぞれの『結合度』である（10, 4, 1）、（8, 3, 2）を座標値とする2本の3次元ベクトルとして表されている。

【0077】次いで、単語A類似度計算部107では、概念空間上において各単語A（目的語）を表現しているベクトル相互のなす角度の余弦を数1に基いて計算し、その計算結果としてそれら単語A同士（A1とA2）の類似度を求める。

【数1】

$$\text{類似度}(A1, A2) = \frac{(1 \ 1 \ 1 \times 1 \ 2 \ 2)}{\sqrt{\sum_{j=1}^n i_{1j}^2} \sqrt{\sum_{j=1}^n i_{2j}^2}}$$

i_{ij} : 単語A i の j 番目の単語Bの結合度

n : 次元数

【0078】図6に示された類義語辞書の内容を参照

し、上記の数を用いて、『coffee』と『tea』との類似度を計算した例を数2に示す。

【数2】

$$\text{類似度}(\text{coffee}, \text{tea}) = \frac{(10 \times 8 + 4 \times 3 + 1 \times 2)}{\sqrt{10^2 + 4^2 + 1^2} \times \sqrt{8^2 + 3^2 + 2^2}}$$

【0079】次いで、単語A類似度計算部107では、全ての単語間について類似度を計算し、類義語辞書を構築する。このようにして構築された類義語辞書の一部を図8に示す。同図から明らかなように、この例では、4個の単語（『coffee』、『tea』、『water』、『egg』）について、相互の類似度が数値化されて記憶されている。

【0080】次に、本発明に係る類義語辞書自動構築装

I drink a cup of coffee. ... 文(1)

(動詞)

【0081】前述したように、このとき、文法上の組を形成する関係として、目的語と動詞の組を形成する関係が設定されていると、文(1)に対して形態素解析処理、構文解析処理、単語関係抽出処理が施された結果として、目的語(coffee)と動詞(drink)とが抽出され、これらの単語は概念空間記憶部105内の該当領域に記憶され、同時に、その組についての出現回数のカウントアップが行われる。

【0082】同様の処理を繰り返しつつ、文書集合を構成する全ての文書の全ての文につき、形態素解析処理、構文解析処理、単語関係抽出処理が施された結果とし

置の具体的な他の実施例を添付図面を参照して詳細に説明する。尚、この例にあっても、説明の便宜上、本発明に係る装置を英文で作成された文書集合に適用したが、日本語で作成された文書集合にも適用できることは勿論である。また、この例では、概念空間上の2点間の距離を計算することにより、語相互の類似度が求められている。この例にあっても、文書集合101より抽出された1文が、文(1)で示されるものであると想定する。

(目的語)

て、中間的に得られた概念空間記憶部の内容の一部を図9に示す。尚、この例では、『結合度』の欄には、目的語（例えば、coffee）が出現した状態にて、動詞（例えば、drink）が出現した出現回数そのものが中間的に記憶されており、未だ単語B結合度計算部106による正規化処理は行われていない。

【0083】図8から明らかなように、この例では、着目された単語A（目的語）である『coffee』は、3個の単語B（動詞）である『drink』、『boil』、『buy』と結合されており、それらの『結合度』はそれぞれ中間的な値として『10』、『4』、

『1』とされている。同様にして、着目された単語A（目的語）である『tea』は、3個の単語B（動詞）である『drink』、『boil』、『buy』と結合されており、それらの『結合度』はそれぞれ中間的な値として『8』、『3』、『2』とされている。

【0084】次いで、単語B結合度計算部106による正規化処理が行われると、中間的に記憶された『出現回数』は、それぞれ正規化されて最終的な『結合度』に変換される。『出現回数』を正規化して最終的な『結合度』を求める方法としては、幾つかの方法が考えられる。

【0085】『出現回数』を正規化して最終的な『結合度』を求める方法として、各目的語毎に、各動詞の出現回数を出現回数が最大の動詞の出現回数で割ると言う手法（以下、第1の手法と称する）を採用した場合に得られる、概念空間記憶部105の内容を図10に示す。同図に示されるように、目的語として『coffee』を例にとると、3個の動詞（『drink』、『boil』、『buy』）のそれぞれとの出現回数は（『10』、『4』、『1』）となるため（図9参照）、それらの最大値である『10』により各出現回数（『10』、『4』、『1』）を除することにより、最終的な結合度（『1.000』、『0.400』、『0.100』）が求められている。

【0086】『出現回数』を正規化して最終的な『結合度』を求める方法として、各目的語毎に、各動詞の出現回数を全動詞の出現回数の合計で割ると言う手法（以下、第2の手法と称する）を採用した場合に得られる、概念空間記憶部105の内容を図11に示す。同図に示

されるように、目的語として『coffee』を例にとると、3個の動詞（『drink』、『boil』、『buy』）のそれぞれとの出現回数は（『10』、『4』、『1』）となるため（図9参照）、それらの出現回数の合計値（『10+4+1』）により各出現回数（『10』、『4』、『1』）を除することにより、最終的な結合度（『0.667』、『0.267』、『0.067』）が求められている。

【0087】尚、以上の第1の手法と第2の手法のいずれを採用しても、実用上十分な精度で類義語辞書の構築は可能であるが、計算機の処理速度が問題となるような場合にあっては、第1の手法を採用するのが好ましいと思われる。

【0088】図10又は図11においてその概念が定量化された2つの単語A（目的語）である『coffee』と『tea』とを概念空間上の点として表現した状態を図12に示す。同図に示されるように、2つの単語A（目的語）である『coffee』と『tea』とは、3個の単語B（動詞）である『drink』、『boil』、『buy』をそれぞれ座標軸とし、かつそれぞれの『結合度』である（（『1.000』、『0.400』、『0.100』）、又は（『0.667』、『0.267』、『0.067』）を座標値とする2個の点として表現されている。

【0089】次いで、単語A類似度計算部107では、概念空間上において各単語A（目的語）を表現している点間の距離を数3に基いて計算し、その計算結果としてそれら単語A同士（A1とA2）の類似度を求める。

【数3】

$$\text{類似度}(A1, A2) = 1 - \frac{\sum_{j=1}^n (I_{1j} - I_{2j})^2}{n}$$

I_{ij} : 単語 A_i の j 番目の単語Bの『結合度』

n : 次元数

【0090】図10に示された類義語辞書の内容を参照

して、上記の数を用いて、『coffee』と『tea』との類似度を計算した例を数4に示す。

【数4】

$$\text{類似度}(\text{coffee}, \text{tea}) = \frac{(1-1)^2 + (0.4-0.375)^2 + (0.1-0.25)^2}{3}$$

$$= 1 - 0.023125/3 = 0.99$$

【0091】次いで、単語A類似度計算部107では、全ての単語間について類似度を計算し、類義語辞書を構築する。このようにして構築された類義語辞書の一部を図13に示す。同図から明らかなように、この例では、4個の単語（『coffee』、『tea』、『water』、『egg』）について、相互の類似度が数値化されて記憶されている。

【0092】尚、以上の各実施例では、着目される『語』（A）を『目的語』、『語』（A）と文法上の組を形成する『関係語』（B）を『動詞』として説明を続けてきたが、『語』（A）と『関係語』（B）との組については種々の変更が可能である。（『語』、『関係語』）に関するその他の組み合わせとしては、（『動詞』、『目的語』）、（『被修飾語』、『修飾語』）、

(『修飾語』, 『被修飾語』)、(『述語』, 『主語』)、(『主語』, 『述語』)の組等が挙げられる。

【0093】

【発明の効果】以上の説明で明らかなように、この発明によれば、『語』相互間の類似度生成に好適な、語の概念を定量化するための方法及び装置を提供することができ、この方法及び装置を利用して語相互間の類似度を生成すれば、文書検索等の用途に適する高性能な類義語辞書を自動的に構築することができる。

【図面の簡単な説明】

【図1】この発明の第1の実施形態である類義語辞書自動構築装置のゼネラルフローチャートである。

【図2】この発明の第1の実施形態である類義語辞書自動構築装置に使用される概念空間記憶部の構造を示す図である。

【図3】この発明の第1の実施形態である類義語辞書自動構築装置に使用される単語B結合度計算部の詳細を示すフローチャートである。

【図4】この発明の第1の実施形態である類義語辞書自動構築装置で構築される類義語辞書の構造を示す図である。

【図5】この発明の第2の実施形態である類義語辞書自動構築装置のゼネラルフローチャートである。

【図6】この発明の第1の実施例である類義語辞書自動構築装置に使用される概念空間記憶部の構造を示す図である。

【図7】図6に示されるの概念空間記憶部に格納された『語』の概念を概念空間上のベクトルとして表現した状態を示す図である。

【図8】この発明の第1の実施例である類義語辞書自動構築装置で構築される類義語辞書の構造を示す図であ

る。

【図9】この発明の第2の実施例である類義語辞書自動構築装置に使用される概念空間記憶部の結合度算出に至る処理途中の状態を示す図である。

【図10】この発明の第2の実施例である類義語辞書自動構築装置に使用される概念空間記憶部に、第1の手法を用いて正規化された結合度を記憶した状態を示す図である。

【図11】この発明の第2の実施例である類義語辞書自動構築装置に使用される概念空間記憶部に、第2の手法を用いて正規化された結合度を記憶した状態を示す図である。

【図12】図10に示されるの概念空間記憶部に格納された『語』の概念を概念空間上の点として表現した状態を示す図である。

【図13】この発明の第2の実施例である類義語辞書自動構築装置で構築される類義語辞書の構造を示す図である。

【符号の説明】

101	文書集合
102	形態素解析部
103	構文解析部
104	単語関係抽出部
105	概念空間記憶部
106	単語B結合度計算部
107	単語A類似度計算部
108	類義語辞書
301	単語B i j 結合度計算部
502	単語A統合類似度計算部
503	統合類義語辞書

【図2】

単語A	単語B	結合度

【図4】

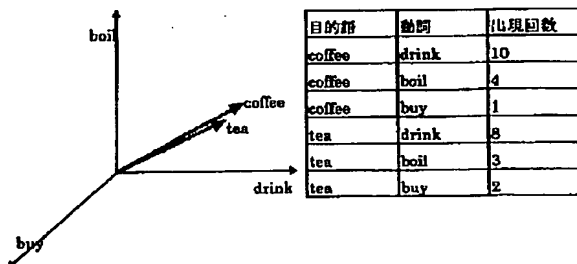
単語	類義語	類似度

【図6】

目的語	動詞	結合度
coffee	drink	10
coffee	boil	4
coffee	buy	1
tea	drink	8
tea	boil	3
tea	buy	2

【図7】

【図9】



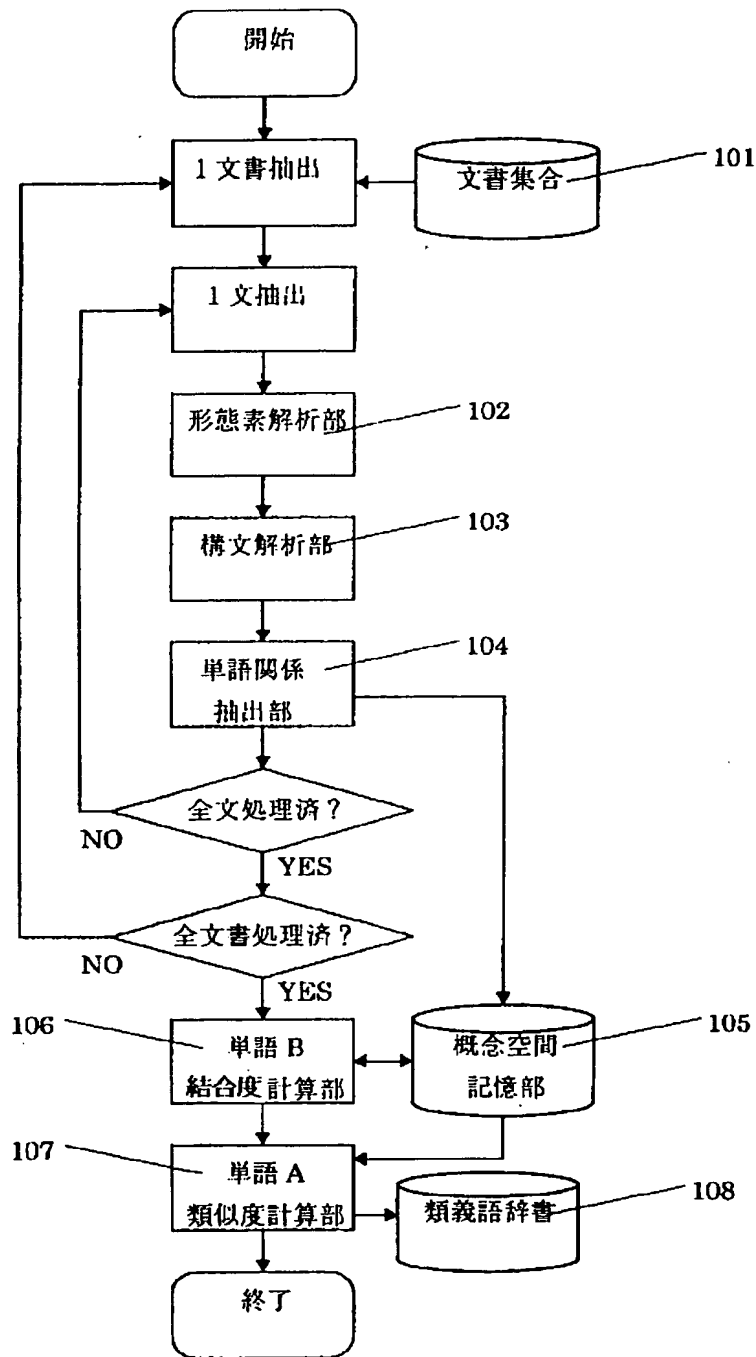
【図10】

目的語	動詞	結合度
coffee	drink	1.000 (10/10)
coffee	boil	0.400 (4/10)
coffee	buy	0.100 (1/10)
tea	drink	1.000 (8/8)
tea	boil	0.375 (3/8)
tea	buy	0.250 (2/8)

【図11】

目的語	動詞	結合度
coffee	drink	0.667 (10/(10+4+1))
coffee	boil	0.267 (4/(10+4+1))
coffee	buy	0.067 (1/(10+4+1))
tea	drink	0.616 (8/(8+3+2))
tea	buy	0.154 (2/(8+3+2))

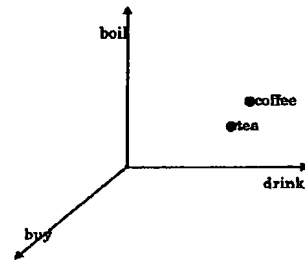
【図1】



【図8】

単語	類似語	類似度
coffee	tea	0.99
coffee	water	0.67
coffee	egg	0.35
coffee
tea	coffee	0.99
tea	water	0.72
tea	egg	0.25
tea
water	coffee	0.67
water	tea	0.72
water	egg	0.13
water
egg	coffee	0.35
egg	tea	0.25
egg	water	0.13
egg
.....

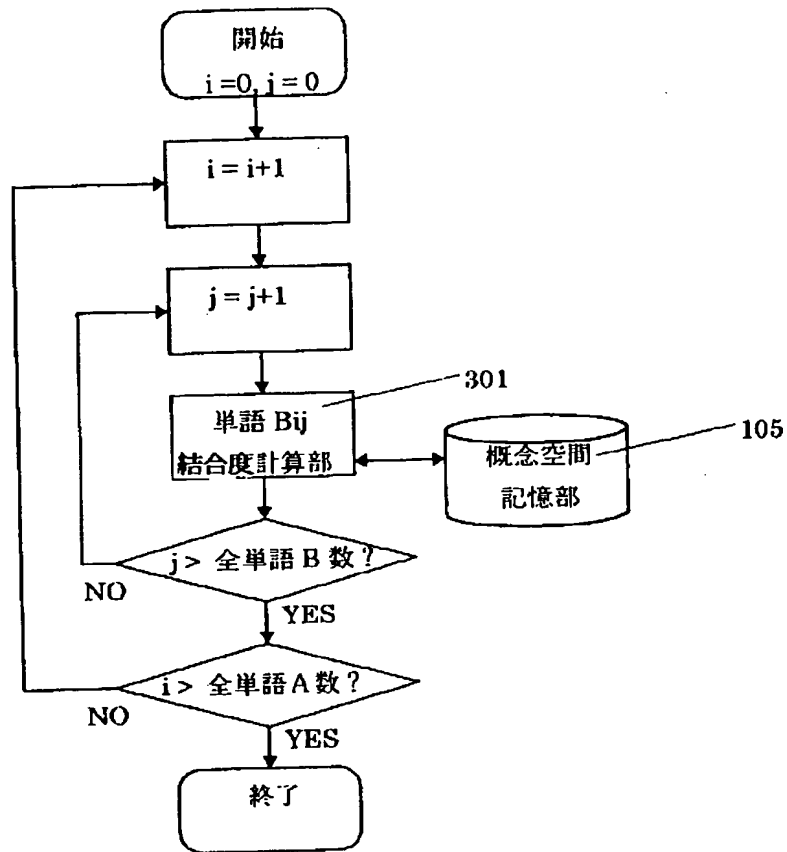
【図12】



【図13】

単語	類似語	類似度
coffee	tea	0.99
coffee	water	0.67
coffee	egg	0.35
coffee
.....

【図3】



【図5】

